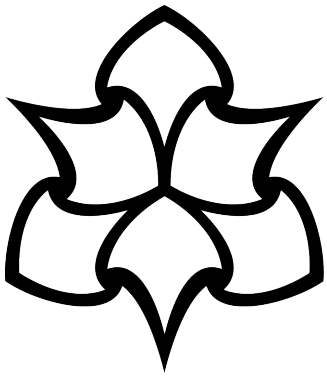
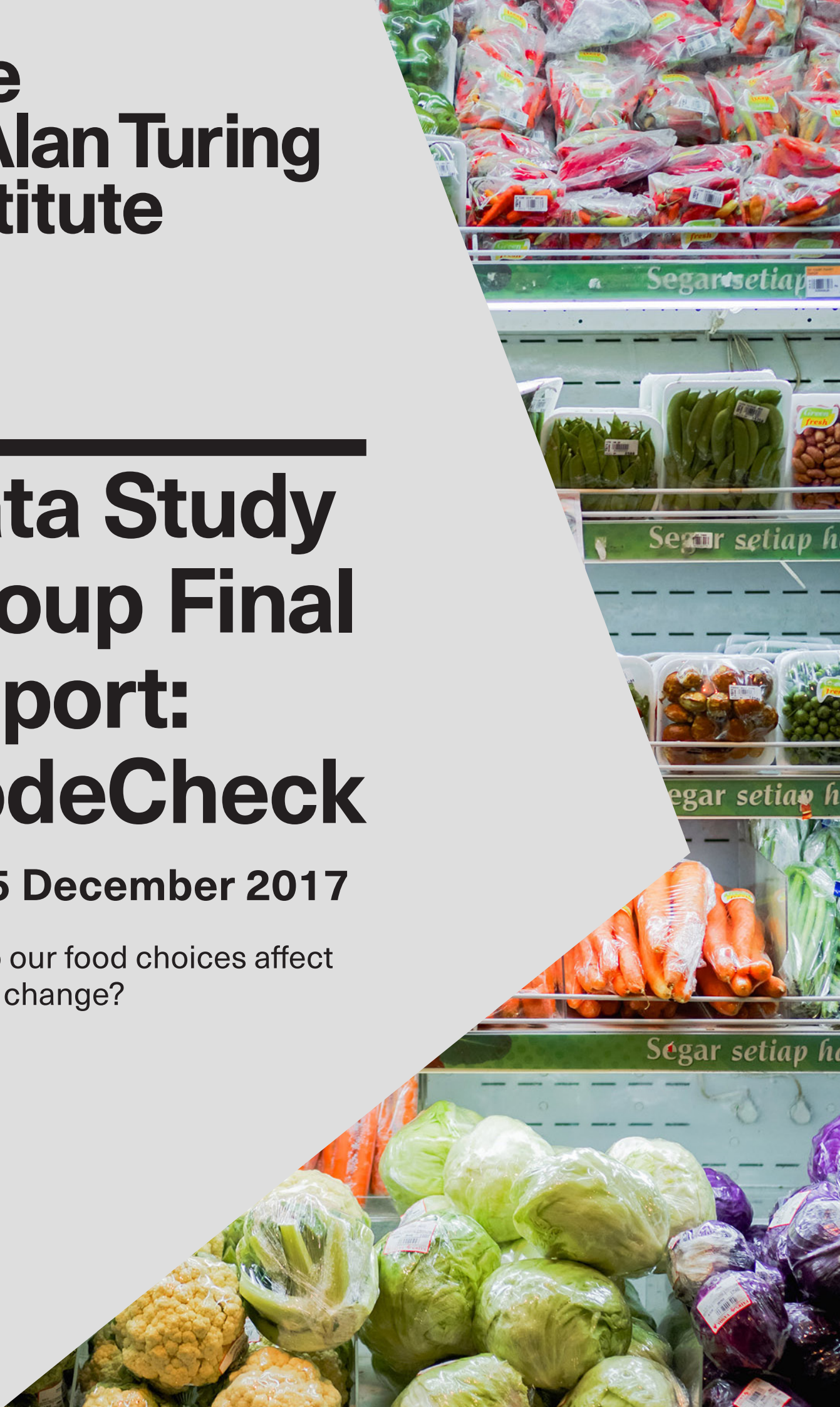


The Alan Turing Institute

Data Study Group Final Report: CodeCheck

11-15 December 2017

How do our food choices affect climate change?



Contents

https://doi.org/10.5281/zenodo.1415344

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

1 Executive Summary

Different approaches were proposed to predict the carbon footprint of products from the different datasets provided by CodeCheck.

Multivariate linear regression and **random forest regression** models perform well in predicting carbon footprint, especially when - in addition to the nutrition information - the product categories, learned through **Latent Dirichlet Allocation (LDA)**, were used as extra features in the models.

The prediction accuracy of the models that were considered varied across datasets.

A potential way to display the footprint estimates in the app was proposed.

1.1 Challenge Overview

Food consumption contributes significantly to global carbon emissions. A potential solution to reduce the carbon emissions associated with food consumption is to provide transparent information about them to consumers. To this end, CodeCheck aims to implement a system whereby information about the carbon footprint of products is presented to the consumers in their app. The hope is that such a system will aid savvy customers to switch to products with lower carbon footprint.

1.2 Main Objectives

The goal for this project was to determine if it is possible to estimate the carbon footprint values for various products. We were given a set of datasets that include expert estimated carbon footprint values for a limited number of products. The aim was to build prediction models on datasets with available carbon footprint values and assess if it is possible to transfer the models to a larger dataset with missing carbon footprint values.

1.3 Overview of the Data

We had access to five datasets of consumer products. Four of which include an estimation of the carbon footprint of the products. The fifth dataset, obtained from the CodeCheck database, is missing estimation of carbon footprint values.

The datasets include information such as the nutritional information and the ingredient make-up the products.

Three of the datasets (O, E, and T) were in English, while the other two were in German (M and Codecheck). Dataset T did not include the ingredient make-up of the products. Detailed information on all the datasets is available in section 3.

1.4 Approach

We considered modelling the CO2 emissions from the products based on their ingredient make-up, their product category, or their nutritional information. For the modelling, we chose simple machine learning models; namely, linear models such as Lasso regression, and tree-based models like Random Forests. Additionally, we used topic modelling to extract informative features from the datasets.

1.5 Main Conclusions

It was possible to build models that can predict the carbon footprints for products based on their ingredient make-up, category and nutritional information, with varying degrees of success depending on the chosen variables.

We showed that there is consistency in the performance of some of the models when transferred to other datasets where the labels are available. However, we were unable to concretely show that these methods can be applied to CodeCheck's database, since we had no way of validating the performance due to lack of labels on this dataset.

1.6 Limitations

One issue with our approach was the difficulty in reusing the models on datasets in other languages. For instance, models trained on dataset E (English) could not be used to make predictions on the CodeCheck dataset (German).

Another consideration was that the estimation methods we developed did not incorporate other sources of carbon emissions, such as those arising by transportation and manufacturing processes.

The estimation methods also failed to incorporate information about the proportion of ingredients in the products, because this information was not consistently available in the datasets.

1.7 Recommendations and further work

It would be very useful to translate the CodeCheck ingredient catalogue into English to aid in the transferability of the models across datasets.

Our results suggested that the carbon footprints values from different datasets are inconsistent (with a discernible offset between various datasets). We suggest that these should be standardised before any further modelling takes place.

Incorporating geolocation data into the models can aid in estimating the carbon footprint due to transportation.

It could be fruitful to consider the change in the behaviour of CodeCheck customers after the introduction of carbon footprint labels. Such analysis can give further insight on the effectiveness of such schemes in combating climate change.

2 Scientific Considerations on Carbon Footprint Estimation Approaches

CodeCheck aims to provide its customers with transparent information about the carbon emissions of products featured in its app. They plan to implement a system that displays the carbon footprint of products to the app users. They hope that by implementing such system their users can make informed decisions about the products they purchase, helping reduce carbon emissions. However, before introducing this feature, certain questions need to be addressed.

2.1 Estimating the Carbon Footprint

Can we estimate the Carbon Footprint of a product based on its ingredient make-up, category and/or nutritional information?

The overall carbon footprint of a product depends on a multitude of factors such as the ingredient make-up, manufacturing process, transportation process, etc. We are given access to datasets containing nutritional information, ingredient make-up and categories of food products. Hence, we chose to focus on these variables to estimate the carbon footprint.

Some of the datasets contain estimates of the carbon footprint produced by each product; however, these are calculated differently for each dataset; therefore, caution needed to be taken when using these estimates in the modelling stage.

It is possible to use the information in the datasets to train models that estimate the carbon footprint of products based on the ingredient make-up. However, given the information that was supplied, it might not possible to obtain a cradle-to-grave estimate of the carbon footprint.

2.2 Estimating the Ingredient Proportions

We were given information about the ingredient proportions for some of the products. One research question that naturally springs up is whether we can use these to estimate the proportion of ingredients that are present in products missing such information?

This is an important consideration when deploying the system in the real-world. Some manufacturers might not include the list of all ingredients present in a product. If ingredient proportion information was shown to be valuable in calculating the carbon footprint for a product, then there would be a need to estimate these proportions in order to come up with more accurate estimates of the carbon footprint.

2.3 Transferring Information Across Datasets

Can we use a model that was trained on one dataset to inform about the carbon footprint of products from a different dataset?

The primary challenge in deploying a system for presenting the carbon footprint to consumers is the lack of such information in CodeCheck's product database. Therefore, it is to transfer the models that are learned on other datasets for use on CodeCheck's own database. This presents many challenges such as, the presence of sample selection bias in the training datasets. Applying transfer learning in this scenario is challenging but nevertheless doable.

3 Data

3.1 Overview

We had accessed to five datasets in total. The datasets consists of records representing products with various details for each such as its nutritional information, country of origin and ingredients (see table 1 for more detailed description of each dataset). In addition, we also had information on generic product categories from CodeCheck's database. For some of the modelling tasks, these were joined to the datasets.

In general, the datasets were in very good quality; hence, in most cases we did not have to perform any data cleaning or preprocessing. In the next section, we highlight cases where some preprocessing was done.

The text data (ingredients) in two of the datasets (CodeCheck and M) were in German, while the rest were in English. This discrepancy made some of the modelling tasks more challenging.

Name	Number of Products	Number of Variables	Carbon Footprint Value	Summary of Variables	Comments
O	162	18	Available	Ingredients, Nutrition, Location	
T	921	7	Available	CF Certification	Omitted from the analysis due to missing information on nutrients and ingredients
E	4153	N/A	Available	Ingredients, Nutrition, Location	Preprocessing was performed on this dataset to obtain the variables and CF values for the products, Product names only available in German
M	254	23	Available	Ingredients, Nutrition, Breakdown of CF	Ingredients in German
CodeCheck	203189	72	Unavailable	Ingredients, Nutrition, Location	Ingredients in German

Table 1: An overview of the available data

3.2 Preprocessing

3.2.1 Text (Ingredient) Data

Data pertaining to the ingredients contained in products was available in text and formatted as a long string for each product. To obtain a more structured representation of this variable, some processing was undertaken.

We followed standard methodology for cleaning text data. This included splitting the strings based on a certain separator (e.g. commas and spaces), removing stop words and obtaining bag-of-word representations for the ingredient list in each dataset.

3.2.2 Dataset E

Some extra preprocessing steps were performed on dataset E. The dataset was available in three parts:

- `E_all_base_products`: containing the names, density and co2 values for "base products", mainly used as ingredients.
- `E_all_combined_products`: containing a list of ingredients included in various "combined products", i.e. products made from base or other combined products.
- `E_nutrient_data`: containing nutritional information for a list of food products, linked to base products.

The two main aims for this preprocessing step are to merge the `co2-value` data from the `E_all_base_products` dataset into the `E_all_combined_products` dataset and to link the nutrient information from the `E_nutrient_data` dataset to the ingredients in the `E_all_base_products` dataset.

The first task allows to obtain CO2 values for the combined products by summing up the CO2 values of its child products. The first step i n this task was standardising the units (to grams). Rows containing units without well-defined gram conversions such as Piece, Pinch and Serving

were excluded. The co2-value field was merged and used to calculate the total CO2, estimated weight based on ingredients and Co2/gram for the combined products.

For the second task, due to the sparsity of the available nutritional data, and to allow easier comparison of nutritional information to dataset O, only columns related to standard mandatory nutrient information (EU regulation 1169/2011) were retained. Units were standardised (g/100g). These were then merged with the E_all_base_products_dataset by nutrition-id.

3.3 Data Quality Issues

3.3.1 Dataset E

There were some issues with the dataset E, particularly with missing nutrient values, spelling errors, and non-unique ids in the E_nutrient_data. Errors the E_all_combined_products for amount of child product in combined product are hard to detect as no total combined product weight is given.

3.3.2 Dataset T

The dataset only contains product names and carbon footprint estimates. Hence, it was not possible to use it in the analysis as it does not contain any features.

3.3.3 Dataset M

There were a number of carbon footprint values that appear to be reported in the wrong units. Additionally, the products are heavily skewed towards chocolate/milk/cheese categories (sample selection bias).

4 Modelling

This section describes the various models and experiments that were attempted during the data study group week. There were three focuses for the modelling stage addressing the scientific considerations presented in Section 2.

The first focus is the prediction of the carbon footprint based on information available in the datasets. For this, we implemented two approaches. The first was constructing informative features from predicting the carbon footprint from the available data, and the second was building models that use these features to predict the footprint. Furthermore, we opted to consider two problems regarding the prediction of the carbon footprint. The first was predicting the CO2 values directly and the second was to predict the level of carbon footprint (low, medium, high).

The second focus was to estimate the proportions of the ingredients present in products. We were unable to get deep insights into this problem; however, we were able to develop a proof-of-concept method that could be used as a blueprint for a more elaborate solution.

The final focus was on assessing the transferability of the models that we created to other datasets. To this end, we made an effort to test some of the models that we developed on various datasets and report some performance measurements when possible.

4.1 Experiment: Regression with Topics as Features

4.1.1 Task Description

Hypothesis and method
We hypothesised that foods that contain similar ingredients might also have similar carbon footprints. Consequently, we opted to use Latent Dirichlet Allocation (LDA) (Blei et. al. 2003) to extract similar groups of foods based on their ingredient lists. LDA is a probabilistic, generative

model that can extract common *topics* between *documents*. In our case, the *documents* are lists of ingredients, and the *topics* are interpreted as groups of 'similar' foods. *Topics* are defined as probability distributions over *words* - each word in the vocabulary is given a conditional probability of appearing in the context of each *topic*. In this instance, *words* are specific ingredients.

As an example, we might infer a *topic* where the following words are likely to appear:

`milk, cocoa, sugar.`

We might interpret this *topic* and call it 'chocolate bars'. If we then see a product with the following ingredient list:

`milk, sugar, cocoa butter, cocoa mass, vegetable fats, emulsifiers, flavourings, cocoa, sugar`

then it will likely have a high probability of belonging to the 'chocolate bar' *topic*. LDA naturally finds groups of similar products by identifying ingredients that tend to appear together.

Once an LDA model has been fit to data, *documents* can be transformed into *document-topic distributions*. A *document-topic distribution* is a vector of probabilities, where there is one entry per *topic*. In our context, these can be interpreted as the probability that a given food product belongs to each of learned food types (*topics*).

We then used the *document-topic distributions* as features for a Lasso regularized linear regression model. In this sense, one can view the LDA step as dimensionality reduction. If the *document topic distribution* is X, then the model is as follows:

$$\log CO_2/gram = \theta X + C$$

so that the logarithm of the carbon footprint is measured as a weighted sum of the components, plus a constant offset C. The components of the vector θ can be interpreted as the contribution of each of the inferred food groups (*topics*) to the overall carbon footprint. For example, we might expect food groups that are likely to contain meat to have a large contribution to the carbon footprint, whereas food groups that are mostly comprised of vegetables might have a small contribution.

We opted for Lasso regularisation because we chose to infer a relatively large number of *topics* (30). It is not obvious that all of these *topics* will be significant i n reality. T he Lasso (L1) regularization allows for some components of θ to be identically zero, so that they are neutral with respect to the carbon footprint. The Lasso will pick out the *topics* that significantly increase or decrease the carbon footprint.

This approach has appeal because it serves two purposes: 1) it automatically groups products into common categories, which may be useful for other reasons, and also provides an estimate of the carbon footprint, 2) it is interpretable: one can observe the common ingredients in the *topics* to identify their meaning. In addition, the coefficients i n the Lasso model are directly interpretable as the contribution of different food groups to the carbon footprint.

Data and preprocessing

We opted to use the datasets E and O for this experiment because both possess ingredient lists in English. Furthermore, both datasets have measured carbon footprints. See Figure 1 below for the distributions of *log CO2* for each of the datasets. Note that there is more dispersion in the *log CO2* distribution of dataset E than in dataset O.

The main data wrangling task was to create a common vocabulary of ingredients across the two databases. In order to maximise the number of shared ingredients across the databases, we split the ingredient lists by word (so that an ingredient called 'flour wheat' would be split into 'wheat' and 'flour') and removed any non-alphabet letter. We then merged the tables to include only ingredients found in both datasets - about a third of the ingredients were found in both sets. For each product, we produced count vectors given the vocabulary. These count vectors are the required input for our modeling approach.

4.1.2 Results

LDA models identify reasonable product categories

LDA produced *topics* that could be easily interpreted as food categories.

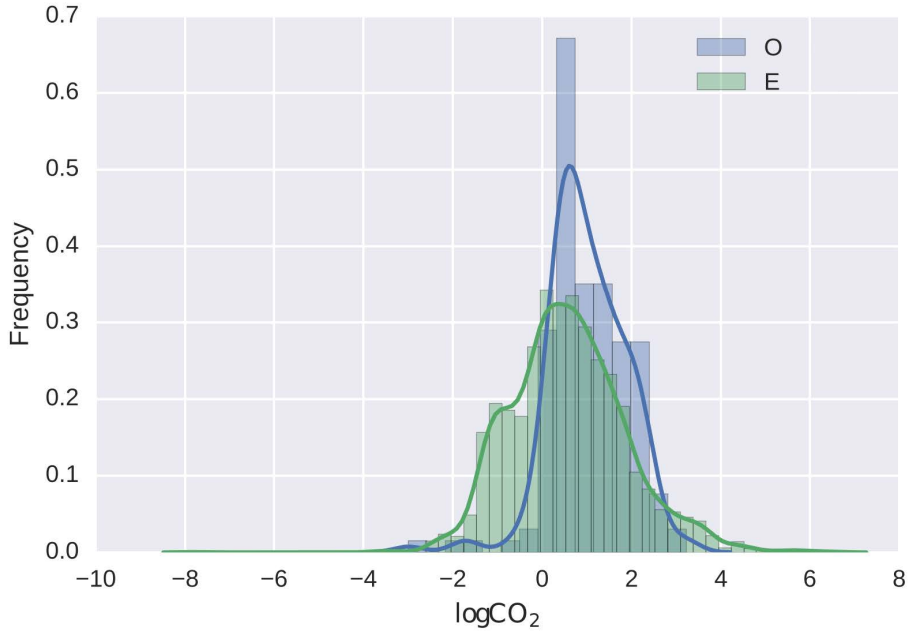


Figure 1: The distribution of log CO2 for Datasets O and E. There is more dispersion in the log CO2 distribution of dataset E than in O.

For instance, one of the categories contained food items such as 'beef', 'spices' and 'vegetables', foods that tend to be in savoury meals. This indicates that the use of LDA for dimensionality reduction is a reasonable choice.

CO2 emissions are correlated with *topics* The linear model trained on dataset E could predict the carbon footprints for dataset O with reasonable accuracy, see Figure 2A . For dataset E, our model produces a **mean squared logarithmic error** (msle) of 1.47 when applied to left out validation data. When the model (now trained on **all** of dataset E) is applied to the dataset O, the msle is 0.92. Note that smaller values of msle indicate more accurate models.

It is important to have a reference point to which we can compare these

msle values. We computed baseline estimates where the mean log carbon footprint of the training set is used as a constant estimate for all products in the test set. For dataset E, this baseline produces an msle of 1.71, so our model is superior. For dataset O, the baseline is 0.75: our model does not beat this when it is trained on dataset E.

We found that using LDA for feature selection was more effective at predicting CO2 emissions than using a simpler Bag-of-Words + linear regression approach, which gave 1.87 msle on dataset O Figure 2B.

As can be seen in Figures 2A and 2B, there is a bias between the predictions for dataset O carbon footprints and the ground truth values. We find that this offset is approximately 0.56. This is perhaps suggestive that there is a systematic difference between the way that the dataset O carbon footprints are measured and dataset E estimates. Removing this bias, the msle becomes 0.61, so that our model beats the baseline. This also highlights a potential application of the model: bringing carbon footprint estimates from different sources onto a common scale (although more investigation is required to determine the source of the bias, as it may well arise from the model).

There was a large cluster of products that were predicted well by the original model without incorporating the offset (Figure 2C (purple)). These products mostly came from the same brand, which largely produced chocolates and sweets. This suggests that either these carbon footprints might be calculated in different ways or there might be a brand-derived bias in carbon emissions, which would suggest that it may be possible to further improve our estimates using brand as a feature.

Lasso model coefficients are consistent with prior knowledge of the environmental effects of food groups

Next, we analysed the coefficients of the Lasso regression model to understand which features were most predictive of carbon footprints. All but two of the thirty LDA features had non-zero coefficients (Figure 2D, so most of the identified food *topics* were predictive of carbon footprint to some degree. The primary food *topics* that were predictive of a low

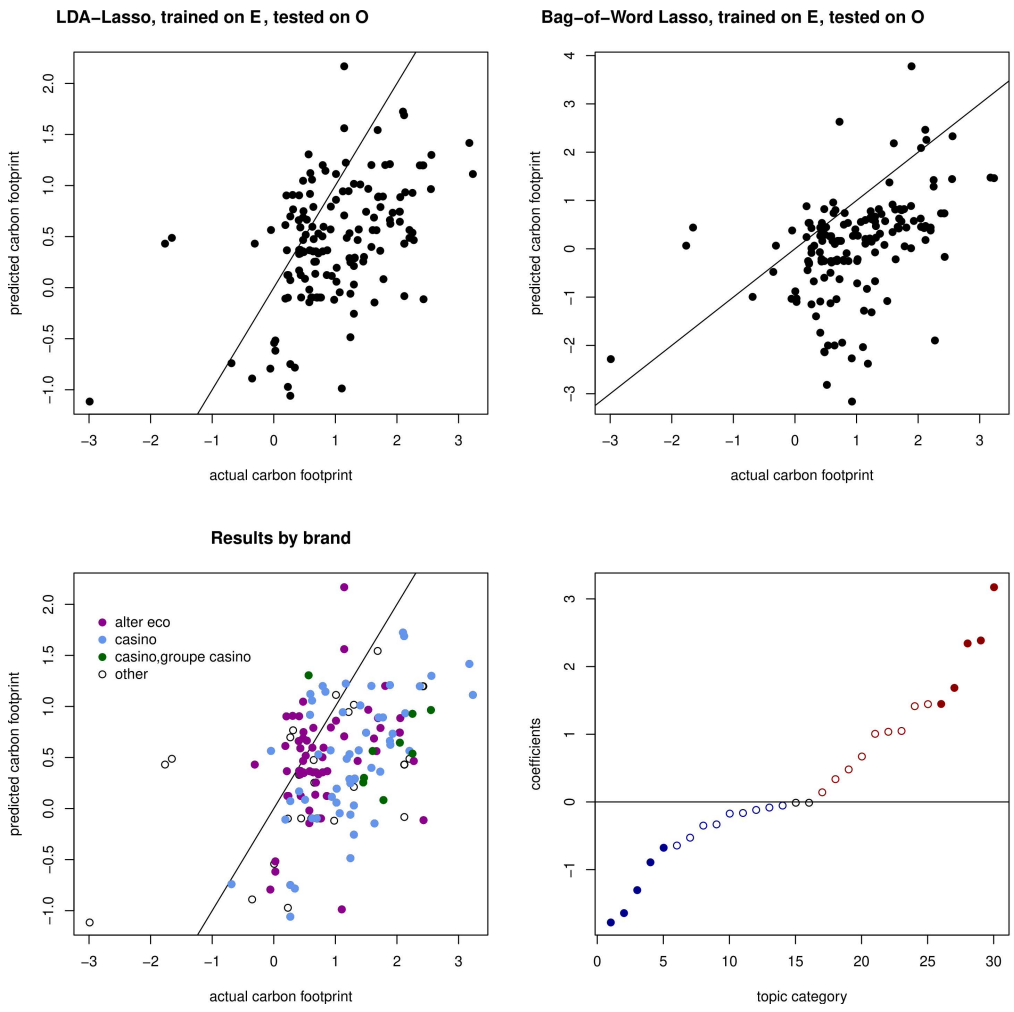


Figure 2: A) LDA-Lasso prediction performance on dataset O, B) BOW-Lasso performance on dataset O, C) LDA-Lasso prediction performance per product cluster, D) Lasso coefficients for LDA topic.

carbon footprint (i.e had large negative coefficients i n t he linear regression model) included fruits, vegetables, and wheat (Figure 3A). The food topics that were linked to a high carbon footprint included beef, cheese, meat, pork, and chicken (Figure 3B). This is consistent with food categories that are known to be good/bad for the environment.

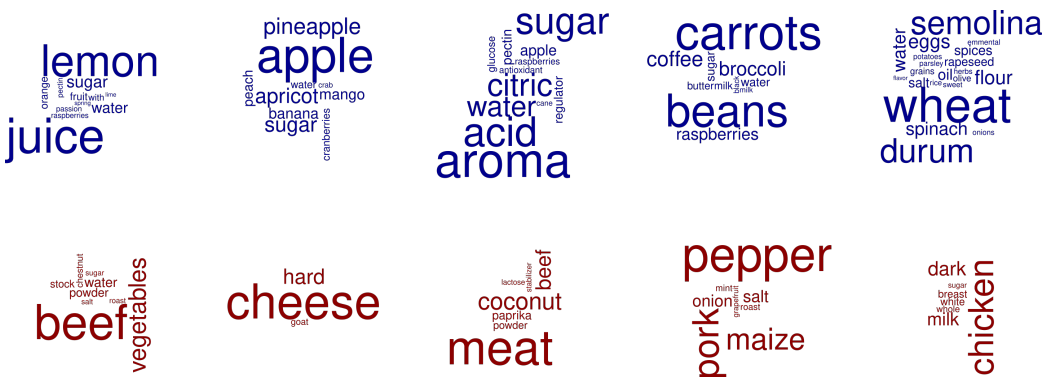


Figure 3: Word cloud representing the A) least (top) and B) most (bottom) LDA topics that are linked to the carbon footprint values.

Performance issues on German-language datasets

We also attempted to test our approach on the German-language dataset M. However, our performance was substantially worse than the other methods that are described below (nutrient-based random forest). This dataset was primarily composed of a small number of categories of products, such as cheese, milk-products and quinoa (Figure 4A), and each of these categories had very similar carbon footprints (Figure 4B). For this reason, we see linear deviations in our linear regression, which represent different category types (Figure 5).

Rank-based approaches do not substantially improve model accuracy

One downside of this approach is that we lose information about the order of the ingredients, which are listed ordered by their proportion in the product. To incorporate this information, we used Kendall's tau to compare the order of ingredients of each product to the ingredient list of each *topic*. More specifically, we calculated Kendall's coefficient between the ordered list of ingredients in each food product, and the ranked list of ingredients for each of the LDA topics. This produces 30 additional features, so that the feature space increases in size to 60 dimensions. However, this did not improve the msle on dataset O(1.03).

Additionally, we attempted to use a Recurrent Neural Network (RNN) to

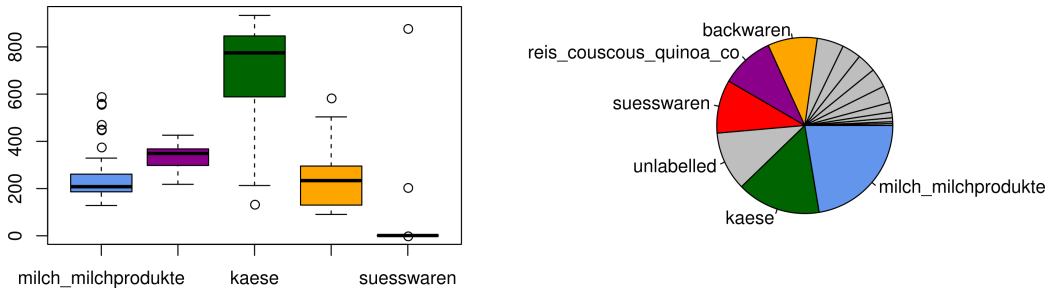


Figure 4: A) Distribution of product categories in dataset M, B) Distribution of carbon footprint per category.

incorporate the ingredient order. We implemented a simple RNN with an embedding layer and linear activation, where the ingredient list is passed as input. We found that this did not outperform the LDA + Lasso model, with an msle of 1.52 on left-out validation data from dataset E.

4.1.3 Limitations

A main difficulty with our approach was that it was difficult to transfer the model to datasets in other languages. This might be particularly challenging for CodeCheck, since their product information is primarily in German, while most complete training set, dataset E, is in English. To resolve this, it would be necessary to consistently translate the ingredients. A translation table of ingredients was available, but incomplete.

In addition, the method does not incorporate other sources of carbon footprint, such as that arising from the transportation or manufacturing processes, unless these factors are correlated with ingredient composition.

We did not incorporate information about the proportion of ingredients in the products, because this information was not consistently available in the datasets; therefore, it was omitted from the analysis.

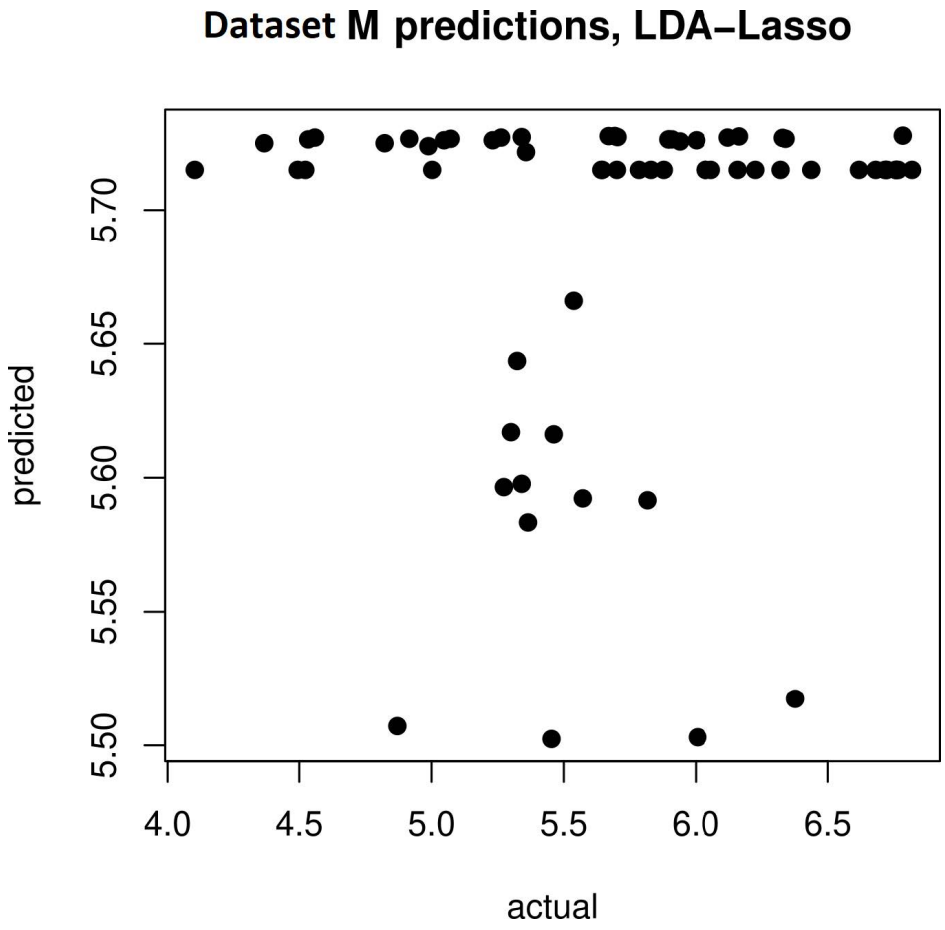


Figure 5: Performance of LDA-Lasso on dataset M.

4.1.4 Take-Home Message

Predicting the carbon footprint
This method was the best method for predicting carbon footprint values in datasets O and E. Moreover, we were effectively able to predict the carbon footprint independent of super-user assigned product categories.

Predicting product type

When a new product is added to the database, it might not have category information. CodeCheck usually relies on super-users to manually add

this information. Using LDA, CodeCheck may be able to automatically detect the product category. To do this, CodeCheck would need to extract the ingredient list for the new product, and identify the probability that it is assigned to each *topic*, forming a vector of probabilities. The product would then be assigned a category according to the *topic* for which it possesses the largest probability.

Carbon footprints are scaled differently across datasets
We noticed a systematic offset between the carbon footprints in the datasets E and O. CodeCheck should consider that each data source might calculate carbon footprint differently, and should therefore be cautious when applying models across datasets.

4.2 Experiment: Modelling the Carbon Footprint on Dataset M

4.2.1 Task Description

The main objective of this experiment is to find a n a ppropriate m odel for the carbon footprint using the ingredient (nutritional) information of the products as well as their categories. The modelling effort is focused on the dataset M. Since the outcome (carbon footprint) is continuous, we primarily considered regression models, specifically linear regression and Random Forest regression. We intentionally chose models with fewer assumptions and less computational challenges, to time and data constraints. It is also desirable to have simpler models as they encourage parsimony.

Data Transformation
For the linear regression model, we applied a log-transformation to the carbon footprint data since its distribution is skewed (see Figure 6).

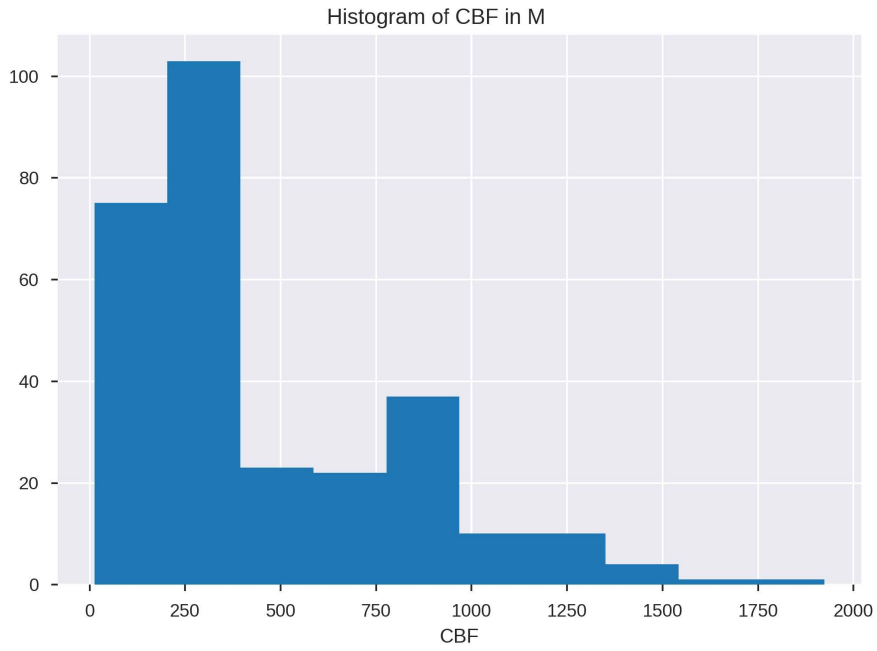


Figure 6: Distribution of the carbon footprint values for the dataset M.

Random Forest Regression
Random Forest regression is a non-linear and non-parametric modelling

technique. It is an ensemble learning method that operates by constructing a multitude of decision trees - a simple machine learning algorithm that models the outputs as leaves of a tree-like graph - at training time and outputting the mean prediction of the individual trees.

Suppose that we have a forest with T trees (i.e., $t = 1, ..., T$). All trees are trained independently and the Random Forest output is the average of all tree outputs as follows (Hastie, et. al. 2001): $Y = \frac{1}{T}Y_t$, where Y_t is the output from the t-th tree.

4.2.2 Results

Multivariate Linear Regression
We first randomly split dataset M into two sets: training (80%) and test (20%). We repeated this splitting 1000 times to ensure the model and its performance is independent of the data splitting process. We fitted the linear regression model, with nutrition information and product categories, to each training set and tested the model on the corresponding test set.

As before, we used the mean square log error as a performance metric, calculating its value for each of the 1000 replications. The average value of this metric was 0.0776.

Figure 7 shows the predicted log carbon footprint against the observed log carbon footprint. The linear regression model performs well when predicting carbon footprint for unseen data.

Random Forest Regression
With the same data splitting and cross-validation processes as above, we fitted a random forest regression model. The average msle for this model was 0.0353.

Figure 8 shows the observed log carbon footprint versus the predicted log carbon footprint values. It suggests that the random forest regression model performs very well (better than linear regression) in predicting the carbon footprint for unseen data on dataset M. This boost in

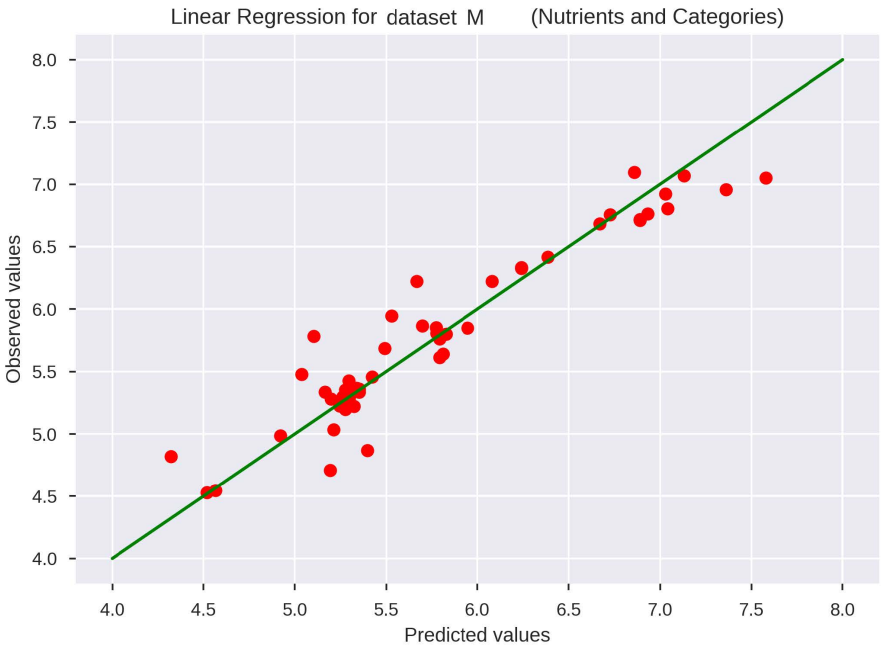


Figure 7: Scatter plot of observed versus fitted values for the linear regression model on dataset M.

Performance is attributed to the random forests' ability to take account non-linear interactions between the inputs, unlike linear regression.

One caveat is that the model trained on the dataset M does not perform well when testing on a different dataset (see the next section). For example, the msle of the random forest regression (trained on dataset M) when testing on the dataset O is 1.3206 which is considered high. Also for this case, when testing on dataset E, the random forest produces a mean square log error of 2.1875 which is very high.

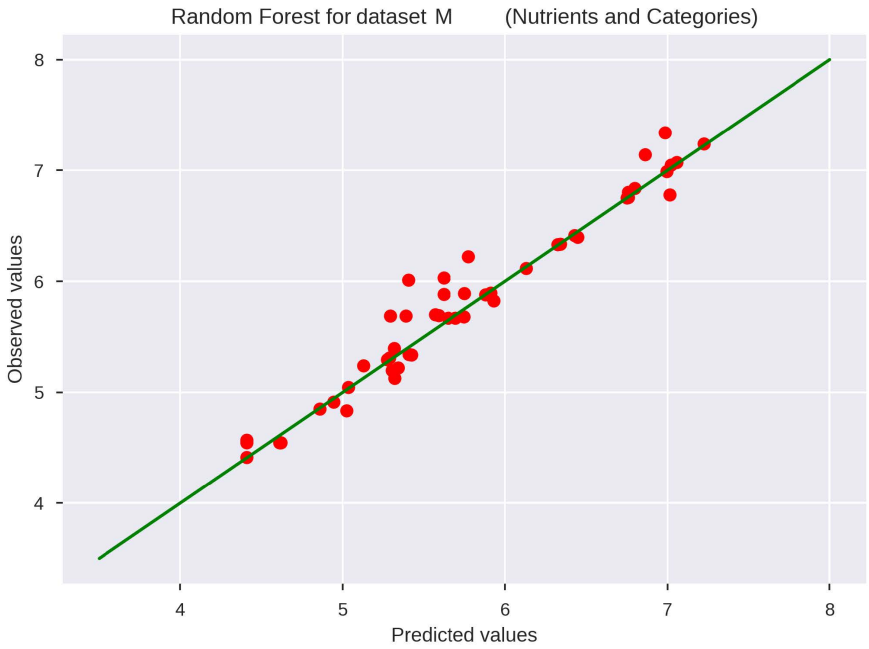


Figure 8: Scatter plot of observed versus fitted values for the random forest regression model on dataset M.

4.2.3 Take-Home Message

For dataset M, both multivariate linear regression and random forest regression performed well in predicting the carbon footprint, especially when, in addition to the nutrition information, the product categories were also added to the model. However, the trained models on dataset M do not perform well when tested on other datasets.

4.3 Experiment: Modelling the Carbon Footprint on Dataset E

4.3.1 Task Description

The objective of this experiment is to use dataset E to train a predictive model for the carbon footprint using the ingredient (nutritional) information. For this experiment, we apply random forest regression.

4.3.2 Results

First, with the same data splitting process as in the previous section, we applied the random forest regression to each training set (80%), and we tested it on each test set (20%). The average of mean square log error using the random forest is 0.7123.

The plot of the observed log carbon footprint values versus the predicted log carbon footprint values is shown in the Figure 9.

Next, we trained the random forest regression on dataset E and then tested the model on the dataset O. The mean square log error was 1.130 which indicates that the predictive performance of the model on the new dataset was not very good. This is evident in the Figure 10.

4.3.3 Take-Home Message

There was a discrepancy on the predictive performance of models trained on one dataset and tested on a different one. A possible reason for this is that the distribution of the ingredients might be different across different datasets. A possible solution would be to integrate different datasets in the training phase of the model.

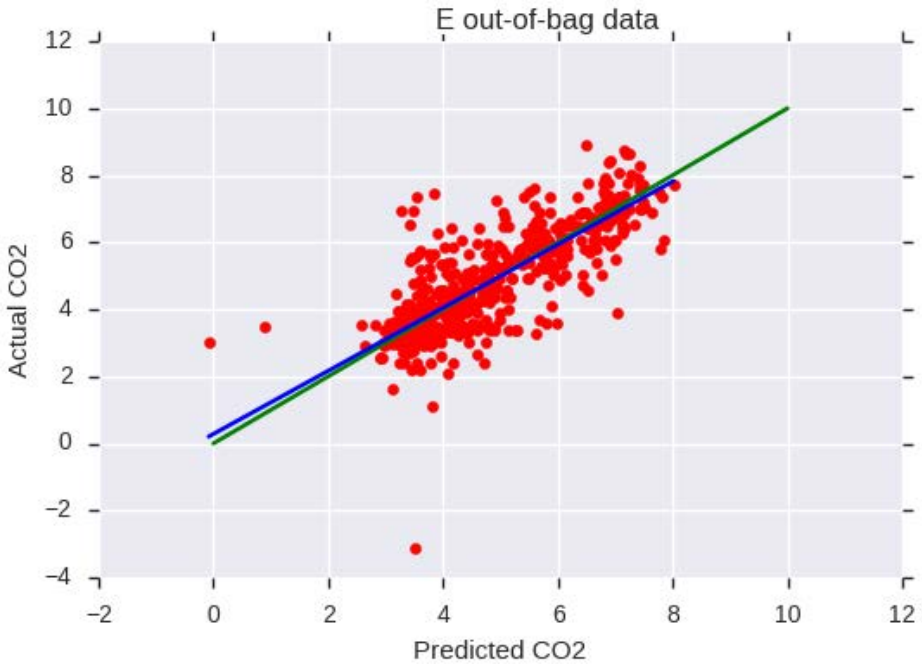


Figure 9: Actual versus predicted log CO2 on dataset E. Blue line represents the fit. Green line represents $x = y$.

4.4 Experiment: Classification with Multi-class Logistic Regression Based on Nutrient Information

4.4.1 Task Description

In this experiment, the goal is to classify products based on their carbon footprint levels. We used the nutritional information of products as predictors. Each product is assigned to one of the following categories: high, medium, low. The task is to predict the class of each product given its nutrition information.

To find 2 reasonable threshold values to use for labeling products with high, medium and low, we visualized the histogram of carbon footprints.

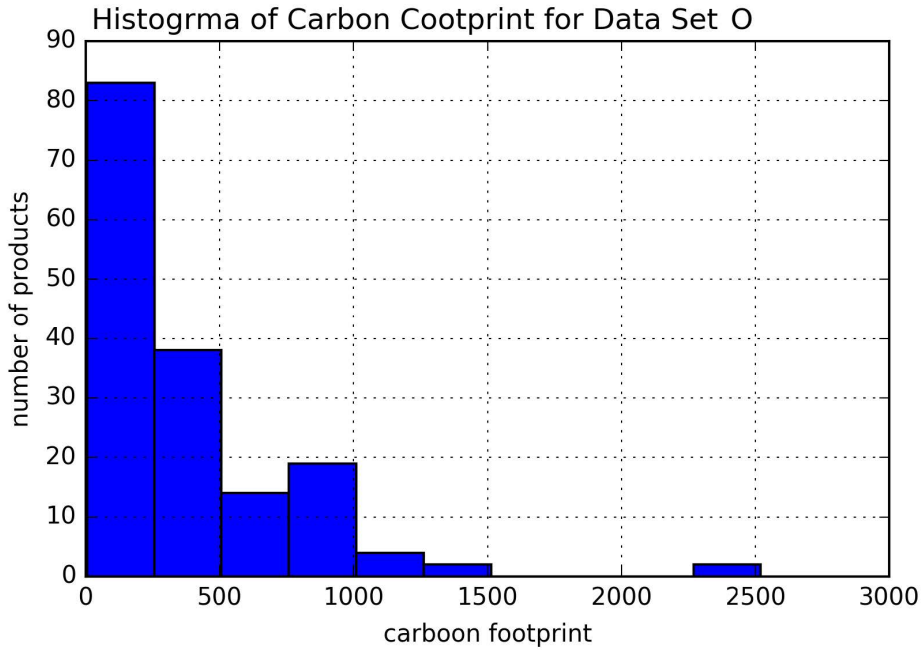


Figure 10: Distribution of the carbon footprint values in dataset O.

We chose 509 and 1000 as cut off points, so that values in ranges 0-509, 509-1000 and 1000-∞ are considered low, medium and high respectively.

The model is trained on dataset O where we split the dataset into training and test sets with proportions 20% and 80%, respectively. Then, we predicted the carbon footprint for the test dataset extracted from O. Furthermore, we tested the trained model on the dataset M.

4.4.2 Results

Figure 12 shows the distribution of the predictions of the model for the carbon footprint level on both dataset O (left) and M (right). As is evident from the figure, the model was able to differentiate between the three assigned levels consistently. Quartiles 1 and 2 have a large proportion of products that are predicted to have a low footprint, while quartiles 4 and 5 have a large proportion predicted as high.

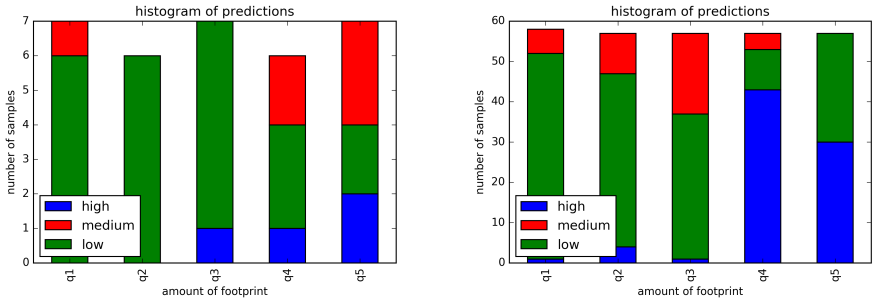


Figure 11: Distribution of the carbon footprint level predictions.
Left: Dataset O, right: M.

4.4.3 Take-Home Message

Using 3 categories (low, medium and high) for representing the carbon footprint is useful for both predictive modelling and user experience. The predictions on the CodeCheck dataset might also be reasonably accurate. However, further investigation is needed in order to verify this hypothesis. In this respect, one might want to analyse the nutrition information of products to see whether the classification labels are reasonable.

5 Future Work and Research Avenues

5.1 Transferring Models Across Datasets

During most of the time spent on this challenge, we focused on building models for predicting the carbon footprint using the labelled datasets. However, the main issue that CodeCheck faces is the lack of labels (carbon footprint values) in their database. One can potentially use Transfer Learning methods to train models on datasets where this information is available and use the trained models to predict the labels for the CodeCheck Database.

A straightforward way to implement this is to use the models that have already been developed to predict on the CodeCheck data. This approach can be benchmarked by measuring the correlation between the predicted outputs of different models. A strong positive correlation indicates that the different models are able to capture some underlying process.

There are two issues to note here. First is that the carbon footprint estimates differ across datasets according to the methodology adopted to calculate this CO2 emission values. This renders any transfer learning approach fruitless unless such inconsistencies are standardised across the source datasets. Another issue is the possibility of a covariate shift between the source and the target datasets. One needs to correct for such phenomenon if it exists before applying any form of transfer.

5.2 Estimating the Contribution of Transportation to the Carbon Footprint

A next step in the analysis is to estimate the contribution to the carbon footprint value due to transporting the product to the location it is bought in. Assuming CodeCheck has the suitable infrastructure, a solution would be to record the geo-location where the product is scanned and then estimate the distance the product travels from its manufacturing location to its scanning location. The distance travelled can be used to estimate

the CO2 emissions by using average figures of CO2/km travelled for various modes of transport.

For instance, the carbon footprint due to transportation for a product produced in Bern, Switzerland and scanned in Berlin, Germany can be estimated by the average amount of CO2/km that a truck emits when travelling from Bern to Berlin. This figure can also take into account EU environmental regulations on the land transportation of products of similar type.

5.3 Presentation of the footprint to the consumer

The following example suggests how information about the carbon footprint values might be presented to the user of the CodeCheck apps.

We assume that consumers are primarily interested in comparing different brands within the same category instead of comparing products in different categories or finding out the average footprint of a product category. Concise labels on products indicating nutrition content are effective tools for persuading consumers to switch between products (Kiesel and Villas-Boas 2013). By providing easily understandable tools for this within category comparison, CodeCheck could also gain competitive advantage over its rivals who tend to display row numbers of the emission that are hard to be interpreted by the layperson.

Research have shown that people want to know how their efforts towards sustainability compare with the efforts of their peers (PAS 2008). Therefore, a desirable labeling design is one that not only shows how a particular product compares to its alternatives in terms of carbon footprint, but also shows the footprint of the typical product other consumers tend to buy.

One suggestion is for CodeCheck to display the footprint in units that are equivalent to the CO2 emission of driving one mile, instead of displaying raw emission rates. Another suggestion is to calculate and present the carbon footprint for 100g of the product instead of the total weight in order to avoid bias from different package sizes.

5.4 Consumer Response Labelling

In order to measure how the labels of the carbon footprint affect purchase habits, research on the behavior on CodeCheck users would be valuable. This might be especially beneficial if we could compare the behaviour consumers who used the app just before the introduction of the carbon footprint labels with their behaviour right after the launch of the new labels.

The data collected by the CodeCheck app does not include which items the consumer purchased – if any – we only observe which items were searched for. Purchase data are property of the retailers where users of CodeCheck shop and these retailers might refrain from sharing their data with any company or research institute. However, recent research identified the factors (e.g. price, brand, package size, ratings) that drive consumer purchase decisions, and the relative magnitude of these factors from product views only (Kim, Albuquerque, Bronnenberg 2010).

A representative survey of CodeCheck consumers before and after the introduction of carbon labels could also help to establish the link between search and purchase behavior, which can then be incorporated into a model that predicts purchase from data on product views.

Information about whether users browse for products having smaller carbon footprints could also be taken into account when deciding about which alternatives (and/or complements) the app should recommend. These recommendation can be personalized as well. For instance, consumers who are more likely to be sensitive to carbon emissions (e.g. based on previous search history) could receive different recommendations than a typical user.

5.5 Beyond the Labels

Food sits at the intersection of environmental, nutritional and health concerns, and is both a cause and a consequence of some of the most pressing challenges we face today (Lang et al. 2009). The production, distribution and delivery of food generate substantial environmental costs. Worldwide agricultural activity, in particular industrialised agriculture and

livestock production, accounts for about a fifth of total greenhouse-gas emissions (McMichael et al. 2007). It is the leading cause of deforestation and biodiversity loss (Tscharntke et al. 2012) and accounts for 70% of all human water usage (UN Water n.d).

Climatic and environmental changes also impact negatively on food production, endangering future food security. In addition, 30–50% of all food produced is spoiled or wasted – representing a waste of land, water and other inputs, 'unnecessary' emissions, and contributing to food insecurity (UN Water n.d.a; Gunders 2012; Parfitt, Barthel and Macnaughton 2010).

Given the complexity of the 'food problem' and the dynamic interplays of cause and effect, simple technical solutions may not sufficient. In this context, perhaps we can consider to extend the original question from "How do our food choices affect climate change?" to "How do whole food systems affect climate change?". Perhaps it is also worth initiating a conversation on what 'critical' role, if any, "data science" could or should play in addressing this question.

References

[1] The economics of climate change mitigation: Policies and options for global action beyond 2012 - oecd. http://www.oecd.org/env/cc/theeconomicsofclimatechangemitigationpoliciesandoptionsforglobalactionbeyond2012.htm. (Accessed on 02/01/2018).

[2] Eu regulation (1169/2011). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2000L0013:20090807:EN:PDF. (Accessed on 02/01/2018).

[3] Greenhouse gas emissions from a typical passenger vehicle — green vehicle guide — us epa. https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle. (Accessed on 02/01/2018).

[4] Guide to pas 2050: How to assess the carbon footprint of goods and services. https://aggie-horticulture.tamu.edu/faculty/hall/publications/PAS2050_Guide.pdf. (Accessed on 02/01/2018).

[5] Regulation (eu) no 1169/2011 of the european parliament and of the council of 25 october 2011 on the provision of food information to consumers, amending regulations (ec) no 1924/2006 and (ec) no 1925/2006 of the european parliament and of the council, and repealing commission directive 87/250/eec, council directive 90/496/eec, commission directive 1999/10/ec, directive 2000/13/ec of the european parliament and of the council, commission directives 2002/67/ec and 2008/5/ec and commission regulation (ec) no 608/2004text with eea relevance. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R1169&from=EN. (Accessed on 02/01/2018).

[6] Water and food security — international decade for action 'water for life' 2005-2015. http://www.un.org/waterforlifedecade/food_security.shtml. (Accessed on 02/01/2018).

[7] Water, food and energy — un-water. http://www.unwater.org/water-facts/water-food-and-energy/. (Accessed on 02/01/2018).

[8] S. Athey and G. W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[11] A. Ghose, P. G. Ipeirotis, and B. Li. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science*, 60(7):1632–1654, 2014.

[12] D. Gunders et al. Wasted: How america is losing up to 40 percent of its food from farm to fork to landfill. *Natural Resources Defense Council*, 26, 2012.

[13] K. Kiesel and S. B. Villas-Boas. Can information costs affect consumer choice? nutritional labels in a supermarket experiment. *International Journal of Industrial Organization*, 31(2):153–163, 2013.

[14] J. B. Kim, P. Albuquerque, and B. J. Bronnenberg. Online demand under limited consumer search. *Marketing science*, 29(6):1001–1023, 2010.

[15] T. Lang, D. Barling, and M. Caraher. *Food policy: integrating health, environment and society*. OUP Oxford, 2009.

[16] A. J. McMichael, J. W. Powles, C. D. Butler, and R. Uauy. Food, livestock production, energy, climate change, and health. *The lancet*, 370(9594):1253–1263, 2007.

[17] J. Parfitt, M. Barthel, and S. Macnaughton. Food waste within food supply chains: quantification and potential for change to 2050. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1554):3065–3081, 2010.

[18] T. Tscharntke, Y. Clough, T. C. Wanger, L. Jackson, I. Motzke, I. Perfecto, J. Vandermeer, and A. Whitbread. Global food security, biodiversity conservation and the future of agricultural intensification. *Biological conservation*, 151(1):53–59, 2012.

[19] S. E. West, A. Owen, K. Axelsson, and C. D. West. Evaluating the use of a carbon footprint calculator: Communicating impacts of consumption at household level and exploring mitigation options. *Journal of Industrial Ecology*, 20(3):396–409, 2016.

[20] L. Whitmarsh. Behavioural responses to climate change: Asymmetry of intentions and impacts. *Journal of environmental psychology*, 29(1):13–23, 2009.

A Team members

- **Angus Williams** is a Data Scientist at the Alan Turing Institute. He applies novel techniques developed at the Turing to real world datasets and problems.
- **Ayman Boustati**. Ayman is a PhD student at the University of Warwick and the Alan Turing Institute. He works on Multitask and Transfer Learning methods for Gaussian Processes. He is the facilitator for this project.
- **Daphne Ezer** Daphne is a Research Fellow at the Alan Turing Institute and the University of Warwick Statistics Department. She conducts research in applications of data science to plant biology.
- **Diego Arenas**. Diego is an EngD in Computer Science student at the University of St Andrews, Scotland, and part of the Analytics team at Aggreko in Glasgow. He works in Data Science and Big Data projects and also enjoys working in Data for Good projects
- **Jan-Hendrik de Wiljes** Jan is a post doc at the University of Hildesheim (Germany). His previous research mainly focused on graphs and hypergraphs defined via number theoretic functions and on cryptology.
- **Marina Chang**
- **Marton Varga**. Marton is a PhD student at INSEAD. In his dissertation he fits econometric models to explain how consumers search for and purchase products online. He also studies the food and the health-care markets.
- **Matthew Groves**. Matthew is a PhD student at the University of Warwick. His research focus is on optimizing information collection methods and their application to machine learning.
- **Reza Drikvandi** Reza is a Research Associate at Imperial College London. His research is mainly focused on statistical modelling for longitudinal and multilevel data analysis, as well as on statistical inference for high dimensional data.
- **Taha Ceritli** Taha is a first year PhD student at the University of

Edinburgh and the Alan Turing Institute. He works on machine learning and automated data science.

turing.ac.uk
@turinginst